DOCUMENT RESUME

ED 190 618 TM 800 412

AUTHOR Bolden, Bernadine J.: Stoddard, Ann
TITLE The Effects of Language on Test Performance of Elementary School Children.
PUB DATE Apr 80
NOTE 25p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level: Elementary Education: Kernel Sentences: *Multiple Choice Tests: Readability Formulas: *Reading Tests: *Sentence Structure: Sex Differences: Test Bias: Test Construction: *Test Format: *Test Interpretation: Test Items

ABSTRACT

This study examined the effect of two ways of question phrasing as related to three styles of expository writing on the test performance of elementary school children. Multiple-choice questions were developed for sets of passages which were written using three different syntactic structures: and which had different levels of difficulty. The questions were phrased using the complete and incomplete stem mode. A total of six different tests were compiled based on stem mode and syntax of the passages. These tests were randomly given to children in two schools (grades two, three, and four). No significant difference in performance was found for sex, school, or grade level for the different syntactic structures. A significant difference in performance was found for the complete versus the incomplete stem mode for the multiple-choice items. Performance was superior on the items with a complete stem; this was true for both schools, both sexes, and for all three grade levels. As might be expected, performance generally increased with grade level. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190618

The Effects of Language on Test Performance of Elementary School Children

Bernadine J. Bolden
Ann Stoddard

JUL 1980
RECEIVED
ERIC/TM

This study examined the effect two ways of question phrasing as related to three styles of expository writing had on the performance of elementary school children. Multiple-choice questions were developed for sets of passages written using three different syntactic structures. The questions were phrased using the complete and incomplete stem mode. A total of six different tests were compiled based on stem mode and syntax of the passages. These tests were randomly given to children in grade levels two, three, and four. No significant difference in performance was found for sex, school, or grade level for the different syntactic structures. A significant difference in performance was found for the complete versus the incomplete stem mode for the multiple-choice items.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
B. Bolden

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 800412

THE EFFECTS OF LANGUAGE ON TEST PERFORMANCE OF ELEMENTARY SCHOOL CHILDREN

Bernadine J. Bolden
Ann Stoddard

Testing is a way of life in our nation's elementary and secondary schools. Not only do teachers periodically administer their own unique tests to their classes, but they also administer the various standardized tests mandated by their local schools or school districts. It has been estimated that more than 250 million standardized tests are given each year (Laosa, 1977), an impressive number when it is realized that there are about 50,000 to 60,000 children between the ages of 5 to 17 years. A recent Newsweek article (Feb. 18, 1980) places the number of standardized achievement and intelligence tests given annually at more than 100 million. Couple this information with the Coleman survey of 1966 which reported that over 90% of the nation's pupils were in schools where IQ and achievement tests were given at the elementary and secondary levels (Kirkland, 1971), and a conclusion can be drawn that school children are tested extensively.

The national impetus toward accountability of educators and of learners seems to be providing an acceleration of the testing phenomenon. Many school systems have instituted or are in the process of instituting policies whereby children will have to perform up to a specified criterion on a designated standardized test before being promoted to the next grade level. Minimum levels of performance on locally developed tests are also being established as graduation requirements in school districts throughout the nation. The entire issue of the May 1978 Phi

Delta Kappan was devoted to a discussion of the issues, implementation problems, and implications surrounding the rise of minimum competency testing. It seems that test scores themselves are increasingly becoming the arbitrators in the educational area.

Tests are used to assess students' progress, identify special needs, and group within schools and within classrooms, among other things. Controversy has existed for decades over the issue of whether tests are fair and unbiased to some groups of children when used in making these and other decisions. The Council on Human Relations of the NEA and the NAACP recommended that a moratorium be placed on testing until some of the issues were resolved. Statistical models designed to eliminate bias in selection have been developed. Numerous studies have been conducted to demonstrate that particular tests do not underpredict for various groups. All of this is in an effort to assess and minimize differences in the effects of the testing phenomenon on various groups within the population.

The issue of test bias is complex and encompasses different areas of concern. Flaugher (1978) lists failure to distinguish between achievement and aptitude tests in test score interpretation, mean group differences, overinterpretation of tests designed to measure particular behaviors, single-group or differential validity, content, selection model used, use of the wrong criterion, and testing atmosphere as examples of areas to which the term test bias is applied. He has also stated that tests themselves cannot be either biased or unbiased, but rather it is in their use that sources of bias occurs (Flaugher, 1974). Newland (1973) echoes this sentiment, and believes that although individual tests may be valid when used with the 70 to 80% of school children who may be classified as "typical," validity decreases when they are used with atypical populations.

Most of the research concerning bias has been in the areas of differential validity and content bias. In addition, the focus has been on bias as it relates to college and job selection of blacks and whites. The results of some of these studies suggest that the tests examined predict as well for blacks as whites, and that content does not significantly affect performance of blacks in an adverse way (Cleary & Hilton, 1968; Flaugher, 1974; Schmeiser & Ferguson, 1978). Schmeiser and Ferguson (1978) further state that although it does not appear that content is a source of bias, other test characteristics may be. Some of these features are item characteristics such as wording and the alternatives of items.

The item format of choice for most standardized tests is the multiple-choice item. Measurement and evaluation texts offer specific suggestions as to how these items should be constructed. Research as to the efficacy of these principles is sparse, however, and test makers have been operating for decades under the assumption that these little-tested principles are valid.

Dunn and Goldstein (1959) stated that they "were unable to uncover any empirical study which specifically investigated the influence of item-writing rules on multiple-choice achievement test difficulty, validity, or reliability" (p. 171). They studied the effect of four construction principles: use of irrelevant cues or specific determiners, the stem as a complete question or incomplete statement, length of alternatives, and grammatical consistency. The population consisted of Army trainees in the 8th week of basic military training. Items containing cues, extra-long correct alternatives, and grammatical inconsistencies were found to be less difficult. Violation of the principles

had no significant differential effect upon reliability or validity.

McMorris, Brown, Snydor and Pruzek (1972) investigated the effects of violating three of the same principles: cues, grammatical inconsistencies, and extra-long correct alternatives. They conducted their study in two suburban schools with eleventh-grade students. Again the violations made the items easier, with no appreciable effect upon reliability and validity.

Another study by Slakter, Koehler, and Hampton (1970) examined the effects cues contained within multiple-choice options, use of obviously incorrect options, and use of options which imply the correctness of each other. Children in grades 5 through 11 were tested. It was found that although items with cues and absurd options seemed easier, mean score differences in response could be explained by random guessing for the item faults until the 8th grade. Specific determiners did not significantly affect results until the 9th grade.

The use of positive and negative stems were tested by Terranova (1969). The subjects consisted of students in middle to upper-class suburban schools in grades 5, 7, 9, and 11. The study was designed to test the effect of positive and negative stems as a function of the frequency of change of direction set (the number of times the subject has to change back and forth from positive to negative stems). The change condition consisted of 0, 1, or 2 changes within the test. Students scored significantly higher on positively stated items than negatively stated items in the 0-change condition. Students also scored significantly higher on the 0- or 1-change condition than the 2-change condition . Thus the difficulty of the test was found to be altered by change in either the stem mode or by change in the frequency of direction

set. No appreciable effect was found on the reliabilities of the tests.

Two studies examined the effect of language usage in the multiple-choice test. The subjects in one study consisted of over 60% black 8th grade and 11th graders, with the mean scores for all subjects near the 33rd percentile on the national norms (Bornstein and Chamberlain, 1970). Subjects in the second study were the entire 8th grade class of a middle class suburban high school (Livingston, 1973). Both studies used the same tests and methods. On one version of the test reading selections, graphs, etc. were unchanged, but the language used in the item stem and alternatives was simplified as much as possible. The second version of the test retained the items in their original form. The results of both studies indicated that the simplified items did not affect achievement to a significant extent.

The available evidence to date does tend to appear to support most of the principles that are accepted as good practice for the construction of multiple-choice items. Even when significant differences did not occur, violation of the principles tended to make faulty items easier than non-faulty items. Two studies appear particularly noteworthy, however. In the study by Slakter, et al. (1970), significant differences in performance attributed to item faults did not appear until 8th grade. Children in the lower grades seemed to respond to something other than the flaw contained within the item. The study by Dunn and Goldstein (1959) did not reflect any significant difference in performance for items using complete statements as the stem or incomplete statements.

All of the studies cited used students in the upper-elementary school grades, junior high, and high school, or adults. There seems to

be a serious problem in the omission of the younger aged elementary school child. It should not be assumed that evidence collected thus far can be generalized to the younger-age group. Boldt, et al. (1977) address this question in an indirect way when stating that the language used in tests is more formal than the written language, and may pose problems for individuals of cultural backgrounds different from the majority middle-class background. It is also possible that the formal language of testing may be reacted to differently by children, according to their age, level of cognitive development, and stage of language development. The study by Slaker, et al. (1970) may suggest just such a phenomenon.

The one study that is most relevant to this investigation (Dunn & Goldstein, 1959) suggests that the use of either a complete or incomplete stem results in non-significant difference in performance. Yet this study included adults only. Texts that delineate principles of constructing the multiple-choice item either do not discuss the two types of stem or state that there is no empirical evidence to support the superiority of one method of writing the stem over the other. The Dunn and Goldstein study has even been cited as evidence to support this view. Two texts (Mehrens & Lehmann, 1973; Sax, 1974) do point out that although there is no evidence to support advantages of either approach, younger children will probably do better with stems that are complete questions. Mehrens and Lehmann (1973) further point out that this approach probably places less demands upon the younger reader.

Reading specialists are quite aware that sentence structure affects readability of written material. It seems reasonable to assume that the structure of the multiple-choice stem cannot be excluded from

considerations of readability of test material. Dawkins (1975) maintains that easy-to-read syntax will keep words together that should go together. It is questionable that this is the case when the stem is presented as an incomplete statement.

Despite the rhetoric and evidence demonstrating the neutral response of students toward complete and incomplete stems of multiple-choice items, Millman (1978) has pointed out that item difficult can be made to fluctuate by changes in how questions are asked. Other related research indicated that questions may be passage independent; that is, correct answers can be derived without reading the passage (Smith & Western, 1978). The implication of this finding is that multiple-choice tests as they are presently constructed raise doubts as to their value.

Children must respond to the prose of passages contained within tests as well as to items measuring different aspects of learning. The syntax of the prose is generally considered important as far as readability is concerned. Coleman (1964, 1965, 1968) demonstrated that with college-age adults readability is improved when transformations are used that have short clauses and active verbs. Readability is also improved by eliminating the use of abstract nouns that have been nominalized from verbs.

Working with elementary school children, Fagen (1971) studied the effects of the following five classes of transformations on reading comprehension: embedding, conjoining, deletion, simple, and position shift. He found that although deletions and embeddings tended to be more difficult to read, they were not the most difficult structures. He also concluded that transformations seem to affect sentence difficulty more than passage difficulty.

Stoodt (1972) examined the effect of using different conjunctions when conjoining sentences. It was found that the easiest conjunctions were and, how, for, and as. Conjunctions such as when, or, but, and if, tended to be more difficult for children to understand. It was also found that there was a relationship between the understanding of the conjunctions used and variables such as sex, socioeconomic level, and intelligence.

Many of the differences in syntactic structure and their possible effect on reading comprehension are not accounted for in the readability formulas which are generally used. According to Klare (1968), word difficulty, usually as a function of frequency, and sentence difficulty, usually as a function of length, are the two most frequently used variables in readability formulas. Botel and Granowsky (1972) claim the "sentence difficulty is the only syntactic measure in readability formulas generally used to control reading difficulty of elementary grade materials ..." (p. 513). These methods of determining readability conflict with some of the views presented by Dawkins (1975). In his analysis of the complexity of syntax he maintains that the omission of connectives such as and, so, when, because, and if, may actually make prose more difficult for children because it violates principles of clear writing. Thus, reading tasks presented in tests which are supposedly controlled for reading level, may in fact pose reading proboems which most currently used reading formulas do not measure.

Purpose

Text bias used in this context means those characteristics of the test which affect individual performance and are not necessarily related to the trait being measured. Specifically, the use of language will be

examined in this study to determine whether the use of specific syntactical structures affect test performance. This will be done by looking at features of the multiple-choice test item and the prose style of test material. Syntax of test items will be limited to choices of two structures: either a complete statement or an incomplete statement in the item stem. Syntax of the prose will consist of passages containing simple sentences only, or passages in which sentences have been combined by using connectives and deletions, or embedding.

Procedures

Three short passages of approximately 120 to 190 words were selected according to two criteria. The first criterion was the probability that the content would be unfamiliar to the children, and thus serve as a measure of their ability to read and comprehend new information. The second was the degree to which the content would appear to be neutral as far as bias toward a particular group. Content contained in the basal reading series used in the school system served as a basis for determining content to which the children had probably been exposed. The selections chosen dealt with interesting facts about three groups of animals (the deep-sea angler, bats, and sharks) which the investigators felt would have the same stimulus appeal for the children in the study. Using the Fry readability formula, one passage was found to be written on the first-grade level, while the two others were both written on a second-grade level.

One passage was written using simple declarative sentences. The other two passages were rewritten using two different techniques. The first style combined sentences and deleted identical sentence elements

to form sentences with relative clauses. For example:

The boy lives near a park. He likes to go walking.

becomes:

The boy who likes to go walking lives near a park.

The second style used conjunction to combine simple sentences while deleting identical elements.

The boy lives near a park and likes to go walking.

Vocabulary remained constant in the process of rewriting, except for the addition of conjunctions, prepositions, and pronouns needed to effect a smooth exposition in the two changed versions.

After rewriting, the newly written passages were analyzed for readability using the Fry formula. The selections were determined to be more difficult in the rewritten versions, although the vocabulary had remained relatively constant. Words used in the rewrite process--such as and, but, that, and which--were determined to be on the first-grade reading level and thus would not in themselves increase the reading difficulty. It was found that one passage increased in difficulty from the first-grade level on the original version, to the second-grade level on the two rewritten versions. The second passage increased from the second-grade level on the original version to grade three on one rewritten version and grade four on the other. The third passage went from the second-grade level on the original version to the fourth-grade level on both rewritten versions. Because of the minor changes in vocabulary, it can be assumed that the differences in readability occurred because of syntactic changes in sentence structure.

Next a series of multiple-choice items were developed for each of the original three passages. For each passage one item emphasized the

main idea and required the selection of an appropriate title while the other items measured literal meaning. Each item was written in two forms, one containing a complete question stated in the stem and the other consisting of an incomplete stem. Options for each item were changed only when necessary for grammatical consistency. Examples of the differences in the finished item are furnished below:

Complete question in stem:

Where does the boy live?

a. near a zoo
b. near a park
c. near a city
c. near a farm

Incomplete stem:

The boy lives near

a. a zoo
b. a park
c. a city
c. a farm

In its final form there existed three versions of the passages in different styles of exposition, and two ways of phrasing questions for the three versions of the passages. The passages were assembled into a test consisting of three different passages written in one of the styles of exposition and one type of item phrasing. There were a total of six different tests, according to the expository style of the passage and the manner of phrasing the stem of the item.

Two elementary schools in Duval County, Florida were chosen in which to administer the tests, one serving children of a predominantly lower SES background, the other serving children of lower to middle SES background. Five second-grade classes, five third-grade classes, and four fourth-grade classes were involved in the study. Grade-level

choices were made because of the assessed reading levels of the selections. Participants were 302 children, comprising 117 second graders, 112 third graders, and 73 fourth graders.

The six tests were given randomly within each classroom. Children in grades two and three were instructed to mark their answers on the test booklet. This was done to eliminate the problem of the tendency of children up to grade three scoring lower on tests on which they use separate answer sheets (Cashen & Ramseyer, 1969). Children in grade four were provided with an answer sheet on which to record their answers. There was no time limit imposed. Before beginning the test, all children read a similar sample selection and responded to items based on it.

To test the hypotheses, the data were analyzed using the factorial design for analysis of variance, multiple-regression approach. The hypotheses tested were:

1. There is no difference in the performance of second-, third-, and fourth-grade children on items with complete or incomplete stem.
2. There is no difference in the performance of second-, third-, and fourth-grade children on passages written in three different expository styles.
3. There is no difference between schools in the performance of children on items with a complete or incomplete stem.
4. There is no difference between schools in the performance of children on passages written in three different expository styles.
5. There is no difference between the sexes in the performance

of children on items with a complete or incomplete stem.

6. There is no difference between the sexes in the performance of children on passages written in three different expository styles.

Results

Analysis of the data indicated a significant difference in performance according to whether a complete or incomplete statement was used in the item stem. (See Table 1.) Lack of significant interaction indicates that this trend holds true across the grade levels tested. Children performed better on items when that item was worded as a complete statement. There was no significant difference in performance on the tests based on different style of syntax used for sentence structure of reading passages. (See Table 2.)

Table 1

Analysis of Variance for Two Stem Modes and Three Grade Levels

	DF	SS	MS	F
Stem Mode	1	87.09	87.09	7.83**
Grade	2	846.82	423.41	38.05**
Interaction	2	7.96	3.98	.36
Within Groups	296	3,305.09	11.13	

**P < .01, N = 302

An analysis of the data according to the two schools also resulted in a significance in performance according to the wording of the item stem (Table 3), but not according to the three syntactic structures of

Table 2

Analysis of Variance for Three Different Syntax Structures and Three Grade Levels

	DF	SS	MS	F
Syntax	2	26.65	13.33	1.19
Grade	2	905.50	452.75	40.35**
Interaction	4	15.86	3.97	.35
Within Groups	293	3,298.95	11.22	

**P < .01, N = 302

Table 3

Analysis of Variance for Two Stem Modes and Two Different Schools

	DF	SS	MS	F
Stem Mode	1	87.09	87.09	6.27*
Schools	1	3.84	3.84	.08
Interaction	1	1.25	1.25	.09
Within Groups	298	4,154.80	13.90	

*P < .05, N = 302

the passages (Table 4). There was no significant difference in performance between schools on either of the variables tested.

Performance on the two variables examined was similar to that according to grade level and school for the two sexes. A significant difference was found in performance according to wording of the item stem (Table 5), but not according to sex. Both males and females responded

Table 4

Analysis of Variance for Three Different Syntax Structures and Two Different Schools

	DF	SS	MS	F
Syntax	2	26.90	13.45	.95
Schools	1	.33	.33	.22
Interaction	2	36.07	18.04	1.28
Within Groups	296	4,183.66	14.09	

Table 5

Analysis of Variance for Two Stem Modes for Males and Females

	DF	SS	MS	F
Stem Mode	1	78.75	78.75	5.77*
Sex	1	5.25	5.25	.38
Interaction	1	.98	.98	.07
Within Groups	298	4,065.52	13.64	

*P < .05, N = 302

better to items worded with a complete stem. Neither sex performed significantly different on the three variations of sentence structures used in the prose material (Table 6).

All analyses resulted in no significant interaction effects, indicating a uniform trend across the variables studied.

Table 6

Analysis of Variance for Three Different Syntax Structures for Males and Females

	DF	SS	MS	F
Syntax	2	31.95	15.97	1.16
Sex	1	2.94	2.94	.21
Interaction	2	16.27	8.14	.59
Within Groups	296	4,095.33	13.84	

Discussion and Conclusion

The results of this study refute the assumption that the use of the complete versus the incomplete stem mode does not affect performance on multiple-choice items. The primary aged children of this study were adversely affected by items utilizing the incomplete statement as the stem. This result was consistent for the two schools sampled, for both males and females, and for the three grade levels tested. The difference in performance between grade levels is to be expected, as the older children have gained in reading proficiency. Therefore this effect is not relevant to the discussion of this study.

It was hypothesized that the children would also show significant differences in performance according to the three types of sentence structures used in the prose selections. Results showed that there was no significant differences. This also was consistent when the data was tested by schools and sex.

An interesting phenomenon can be seen when examining the tables of means for the data, however. The means for the passages written using combining techniques (Table 8) are consistently higher than the

means for passages containing simple sentences. This is true for the three grade levels tested. By examining this same table it can also be seen that the means obtained on passages using the embedding technique are also higher than the means obtained on passages using simple sentences, except for grade four, where they are equal.

Table 7

Mean Test Scores on Two Stem Modes
For Three Grade Levels

Grade Level	Complete Stem N	Complete Stem Mean	Incomplete Stem N	Incomplete Stem Mean
Second	51	8.37	66	7.53
Third	56	11.07	56	10.25
Fourth	42	12.07	31	11.94

Table 8

Mean Test Scores on Three Different Syntax Structures
For Three Grade Levels

Grade Level	Simple N	Simple Mean	Embedding N	Embedding Mean	Combining N	Combining Mean
Second	34	7.53	41	7.76	42	8.33
Third	37	9.81	38	11.11	37	11.05
Fourth	23	11.83	23	11.83	27	12.33

The means for the two schools (Table 10) indicate that in school 2 the children also scored best on the passages containing the combined sentence structures. The data for sexes (Table 12) shows this same effect for females. Dawkins (1975) indicated that when syntax is more readable, children's reading performance increases. However, children

in this study show a consistent trend toward better performance on materials that most readability formulas would suggest is more difficult for a specific reading level.

Table 9

Mean Test Scores on Two Stem Modes
For Two Schools

School	Complete Stem N	Complete Stem Mean	Incomplete Stem N	Incomplete Stem Mean
School A	90	10.30	93	9.43
School B	58	10.76	60	9.40

Table 10

Mean Test Scores on Three Different Syntax Structures
For Two Schools

School	Simple N	Simple Mean	Embedding N	Embedding Mean	Combining N	Combining Mean
School A	56	9.80	58	9190	69	9.87
School B	37	9.16	44	9.95	37	11.11

Table 11

Mean Test Scores on Two Stem Modes
For Males and Females

Sex	Complete Stem N	Complete Stem Mean	Incomplete Stem N	Incomplete Stem Mean
Females	72	10.63	82	9.49
Males	77	10.25	71	9.34

Table 12

Mean Test Scores on Three Dfffferent Syntax Structures For Males and Females

Sex	Simple N	Simple Mean	Embedding N	Embedding Mean	Combining N	Combining Mean
Females	50	9.40	49	9.88	55	10.71
Males	44	9.57	53	9.96	51	9.86

Two potential practical ramifications of this investigation should be noted. First the major finding was that item difficulty as determined by question format, affected student performance. This finding suggests that where elementary children are concerned, extreme care should be given to test construction based on their intellectual development. What is clear from the data of this study is the relatively high probability that with elimination of test bias, young children will show gains on specific proficiencies being measured.

Another conjecture is that test language and question configuration may not be as significant as has been believed in affecting student test performance, especially when passage independent items are considered. This study did not treat this issue; consequently, further study that would contribute toward the establishment of criteria for controlling passage independent items would help to make the interpretation of test results more meaningful.

Finally, since a positive effect for the writing style of the passage and grade levels was found it seems important that measures other than just test scores be used to make educational decisions for young children, such as placement in reading groups. Moreover, additional research is needed to isolate variables which might influence the

effectiveness of syntactic structures in test making and the prose of children's reading selections. Future research focusing on test bias, language usage and question format should provide findings that will prove useful in test construction applications.

REFERENCES

Boldt, R.F., Levin, M.K., Powers, D.E., Griffin, R.C., Troike, R.C., Wolfram, W. and Ratliff, F.R. Sociolinguistic and measurement considerations for construction of armed services selection batteries: Final report for period October 1975-June 1977. Air Force Human Resources Laboratory, Brooks AFB, Texas, 1977. ED 151 405

Bornstein, H. & Chamberlain, K. An investigation of the effects of "verbal load" in achievement tests. American Educational Research Journal, 7(4), 1970, 597-604.

Botel, M. & Granowsky, A. A formula for measuring syntactic complexity: A directional effort. Elementary English, 1972, 49, 513-516.

Breland, H.M. & Stocking, M., Pinchak, B.M., & Abrams, N. The cross-cultural stability of mental test items. ETS, Princeton, N.J., 1974. ED 137 370.

Cashen, V.M. & Ramseyer, G.C. The use of separate answer sheets by primary age children. Journal of Educational Measurement, 1969, 6(3), 155-158.

Cleary, T.A. & Hilton, T.L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Coleman, E.B. The comprehensibility of several grammatical transformations. Journal of Applied Psychology, 1964, 48(3), 186-190.

_____. Learning of prose written in four grammatical transformations. Journal of Applied Psychology, 1965, 49(5), 332-341.

_____. Experimental studies of readability: Part I. Elementary English, 1968, 45, 166-178.

Dawkins, J. Syntax and readability. International Reading Association: Newark, Delaware, 1975.

Dunn, T.F. & Goldstein, L.G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 1959, 19, 171-179.

Eagle, N. & Harris, A.S. Interaction of race and test on reading performance scores. Journal of Educational Measurement, 1969, 6(3), 131-135.

Fagan, W.T. Transformations and comprehension. Reading Teacher, 1971, 25, 169-172.

Flaugher, R.L. Bias in testing: A review and discussion. December, 1974, ETS, TM Report 36.

_____. The many definitions of test bias. American Psychologist, July 1978, 671-679.

Kirkland, M.C. The effects of tests on students and schools. Review of Educational Research, 1971, 41(4), 303-350.

Klare, G.R. The role of word frequency in readability. Elementary English, 1968, 45, 12-22.

Laosa, L.M. Nonbiased assessment of children's abilities: Historical antecedents and current issues. In T. Oakland (Ed.), Psychological and Educational Assessment of Minority Children. New York: Brunner/Mazel, Publishers, 1977.

Livingston, S.A. "Verbal overload" and achievement tests: A replication. American Educational Research Journal, 1973, 10(2), 155-162.

McMorris, R. F., Brown, J.A., Snyder, G.W., & Pruzek, R.M. Effects of violating item construction principles. Journal of Educational Measurement, 1972, 9(4), 287-295.

Mehrens, W.A. & Lehmann, I.J. Measurement and Evaluation in education and psychology. New York: Holt, Rinehart and Winston, Inc., 1973.

Millman, J. Determinants of item difficulty: A preliminary investigation. CSE Report No. 114. July, 1978, ED 163 071.

Newland, T.E. Assumptions underlying psychological testing. Journal of School Psychology, 1973, 11(4), 316-322.

Newsweek, February 18, 1980, 97-104.

Phi Delta Kappan, May, 1978.

Sax, G. Principles of educational measurement. Belmont, California: Wadsworth Publishing Company, Inc., 1974.

Schmeiser, C.B. & Ferguson, R.L. Performance of black and white students on test materials containing content based on black and white cultures. Journal of Educational Measurement, 1978, 15(3), 193-200.

Slakter, M.J., Koehler, R.A., & Hampton, S.H. Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 1970, 7(2), 119-122.

Smith, C.F., Jr., & Western, R.D. Reading proficiency and standardized tests: Another objection. Paper presented at the annual meeting of the National Reading Conference, St. Petersburg, Florida, 1978. ED 165 125.

Stoodt, B.D. The relationship between understanding grammatical conjunctions and reading comprehension. Elementary English, 1972, 49, 502-504.

Terranova, C. The effects of negative stems in multiple-choice test items. Dissertation Abstracts, 1969-70, 30, 2390-A.